 Nucleic Acids Research, 1999, Vol. 27, No. 14 **2825–2831**

Construction of a variability map for eukaryotic large subunit ribosomal RNA

Abdelghani Ben Ali, Jan Wuyts, Rupert De Wachter, Axel Meyer[1] and Yves Van de Peer*

Department of Biochemistry, University of Antwerp (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium and [1]Department of Biology, University of Konstanz, D-78457 Konstanz, Germany

Received May 14, 1999; Accepted May 27, 1999

ABSTRACT

In this paper, we present a variability map of the eukaryotic large subunit ribosomal RNA, showing the distribution of variable and conserved sites in this molecule. The variability of each site in this map is indicated by means of a colored dot. Construction of the variability map was based on the substitution rate calibration (SRC) method, in which the substitution rate of each nucleotide site is computed by looking at the frequency with which sequence pairs differ at that site as a function of their evolutionary distance. Variability maps constructed by this method provide a much more accurate and objective description of site-to-site variability than visual inspection of sequence alignments.

INTRODUCTION

A few years ago, quantitative substitution rate maps of the 5S rRNA, the small subunit ribosomal RNA (SSU rRNA), and the large subunit ribosomal RNA (LSU rRNA) of bacteria (1) and of the SSU rRNA of eukaryotes (2) were published. These maps were constructed by applying the substitution rate calibration (SRC) method that defines the variability or substitution rate of each nucleotide site as its evolutionary rate relative to the average evolutionary rate of all the nucleotide sites of the molecule (3,4). The variability maps constructed in this way clearly showed the distribution of variable and conserved positions in the different rRNA molecules.

Until recently, a reliable variability map for the eukaryotic LSU rRNA could not be constructed because too few complete sequences were available. However, due to recent sequencing efforts, the number of complete eukaryotic LSU rRNA sequences has increased significantly (5,6, unpublished results) and currently about 80 complete sequences are available for organisms belonging to the so-called crown of evolution (7), which now allows the construction of a reliable and detailed map showing the substitution rates in the LSU rRNA of eukaryotes.

Detailed information about the variability or conservation of nucleotide positions in ribosomal RNA is important for several reasons. The variability maps can be interpreted in terms of higher order structure and function and they facilitate the selection of areas suitable for the design of PCR primers and hybridization probes. In addition, the measurement of site variability is important from a phylogenetic point of view. While conserved areas can be used to unravel old relationships, the more variable regions can be used to study evolution between closely related organisms (e.g. 8). Regarding phylogenetic tree construction, the study of site variability in molecules has gained much interest lately. Newly developed tree construction methods take into account differences in nucleotide substitution rates, which leads to more consistent tree topologies (9–11).

MATERIALS AND METHODS

Sequence alignment and estimation of substitution rates

The LSU rRNA database, established at the University of Antwerp (UIA) in 1994, is continuously updated by scanning the international nucleotide sequence libraries such as GenBank and EMBL for corrected or new ribosomal RNA sequences. In general, only complete or nearly complete sequences are compiled. All ribosomal RNA sequences are stored in the form of an alignment that is based on the secondary structure adopted for the molecule (e.g. 12–14). All sequences in the database are aligned with the DCSE sequence editor (15). Beside the primary and secondary structure information, literature references, accession numbers and detailed taxonomic information about the organism from which the sequence was derived are also compiled. For more information about the LSU rRNA database and its contents, we refer to the latest database issue of Nucleic Acids Research (16). The easiest way to obtain data is through the World Wide Web at URL http://rrna.uia.ac.be/lsu/

Estimation of substitution rates and construction of a variability map was done as described previously (1,3,4). In short, substitution rates are estimated by looking at the frequency with which sequence pairs differ at each site as a function of the distance between the sequence pairs (3). Substitution rates or variabilities v are estimated for every site in the sequence alignment that is not absolutely conserved and contains a nucleotide in at least 25% of the aligned sequences. Then, after estimation of all substitution rates, alignment positions are grouped into sets of similar rate. A spectrum of relative nucleotide substitution rates is thus obtained (1,4). A color map, superimposed on the secondary structure model of the LSU rRNA, can then be constructed by dividing the nucleotides into

*To whom correspondence should be addressed at present address: Department of Biology, University of Konstanz, D-78457 Konstanz, Germany. Tel: +49 7531 88 2763; Fax: +49 7531 88 3018; Email: yves.vandepeer@uni-konstanz.de

Table 1. Eukaryotic LSU rRNA sequences (and accession numbers) used for the nucleotide substitution rate calibration

Animals (18)	
Acipenser brevirostrum	U34340
Aedes albopictus	L22060
Anguilla rostrata	U34342
Anopheles albimanus	L78065
Caenorhabditis elegans	X03680
Drosophila melanogaster	M29800
Dugesia tigrina	U78718
Herdmania momus	X53538
Homo sapiens	M11167, J01866
Latimeria chalumnae	U34336
Lepidosiren paradoxa	U34337
Mus musculus	X00525, J00623
Neoceratodus forsteri	U34338
Oncorhynchus mykiss	U34341
Protopterus aethiopicus	U34339
Rattus norvegicus	V01270
Xenopus borealis	X59733
Xenopus laevis	X02995
Fungi (11)	
Arxula adeninivorans	Z50840
Blastocladiella emersonii	X90411, X90410
Candida albicans	X70659, L07796
Cryptococcus neoformans	L14067
Entomophaga aulicae	U35394
Pneumocystis carinii	M86760
Saccharomyces cerevisiae	J01355, K01048
Saccharomycopsis fibuligera	U09238, U10409
Schizosaccharomyces japonicus	Z32848
Schizosaccharomyces pombe	Z19578
Tricholoma matsutake	U62964
Land plants (21)	
Acorus gramineus	AF036490
Arabidopsis thaliana	X52320
Brassica napus	D10840
Drimys winteri	AF036491
Ephedra distachya	AF036489
Eucryphia lucida	AF036494
Fragaria ananassa	X58118, X15589
Funaria hygrometrica	X99331, X74114
Gnetum gnemon	AF036488
Hamamelis virginiana	AF036495
Jepsonia parryi	AF036497
Lithophragma trifoliata	AF036501
Mitella pentandra	AF036502
Oryza sativa	M11585,M16845
Parnassia fimbriata	AF036496
Peltoboykinia tellimoides	AF036499
Plumbago auriculata	AF036492
Saxifraga mertensiana	AF036498
Sinapis alba	X66325
Tellima grandiflora	AF036500
Tragopogon dubius	AF036493
Green algae (1)	
Chlorella ellipsoidea	D17810
Heterokont algae and relatives (8)	
Hyphochytrium catenoides	X80345, X80346
Nannochloropsis salina	Y07975, Y07974
Ochromonas danica	Y07977, Y07976
Phytophthora megasperma	X75631, X75632
Prorocentrum micans	X16108, M14649
Scytosiphon lomentaria	D16558
Skeletonema pseudocostatum	Y11512, Y11511
Tribonema aequale	Y07979, Y07978
Apicomplexans (6)	
Cryptosporidium parvum	AF040725
Eimeria tenella	AF026388
Neospora caninum	AF001946
Plasmodium falciparum	U21939
Theileria parva	AF013419
Toxoplasma gondii	X75429
Ciliates (4)	
Tetrahymena pyriformis	X54004, M10752
Tetrahymena thermophila	X54512
Spathidium amphoriforme	Unpublished
Euplotes aediculatus	Unpublished
Haptophytes (2)	
Phaeocystis antarctica	Unpublished
Prymnesium patelliferum	Unpublished
Red algae (2)	
Gracilaria verrucosa	Y11508, Y11507
Palmaria palmata	Y11506
Other (4)	
Chlorarachnion sp.	Unpublished
Pedinomonas minutissima	U58510
Guillardia theta	Unpublished
Guillardia theta nucleomorph	Y11510, Y11509

Duplicate sequences belonging to the same species were omitted from the analysis. When two accession numbers are given, the first one is that of the 28S rRNA, the second one that of the 5.8S rRNA.

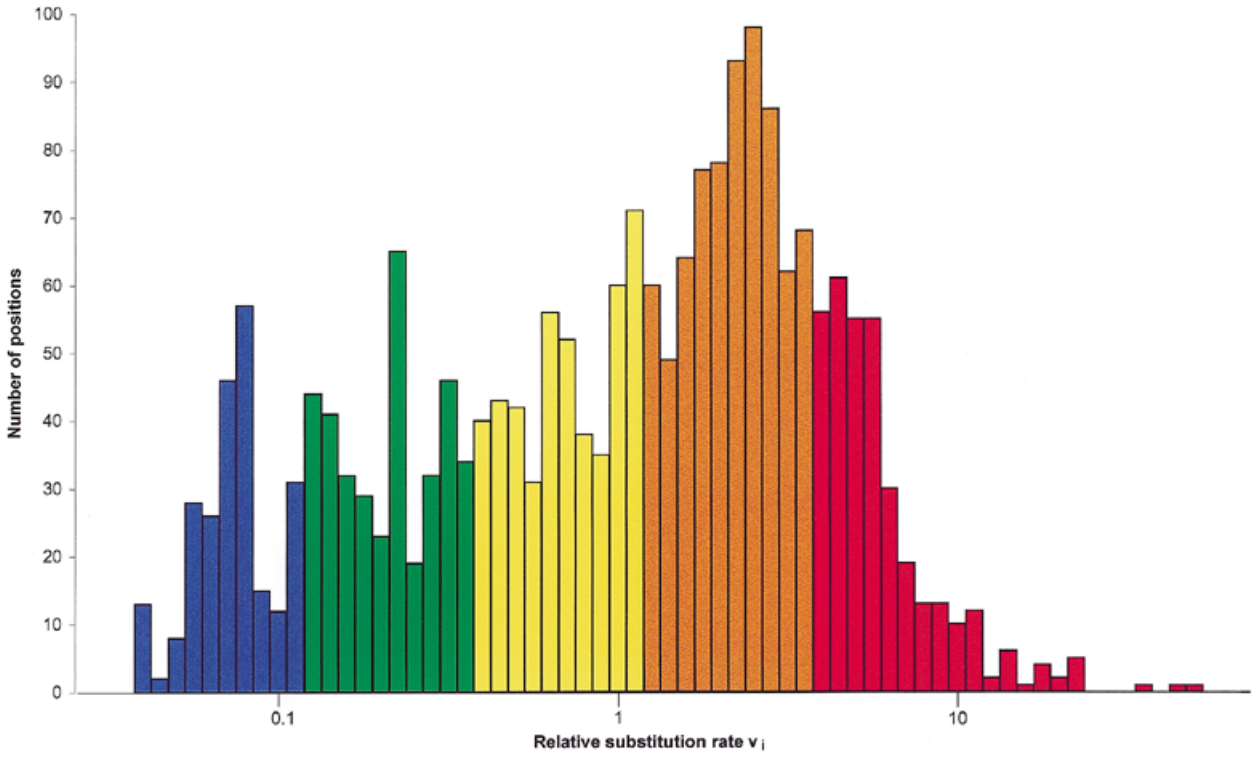


Figure 1. Distribution of the relative substitution rates, estimated from an alignment of 77 eukaryotic LSU rRNAs. The species used are specified in Table 1. Rates were estimated for 2153 alignment positions. Not included are 923 positions that are identical in all known sequences and 474 positions that contain a nucleotide in <25% of the aligned sequences. Sets of nucleotides are indicated in the same colors as used on the variability map of Figure 3.

different variability subsets and assigning a different color to each of these.

Once the shape of the spectrum is known, it is also possible to derive a new equation describing the evolutionary distance between two sequences as a function of the observed number of differences, i.e. the dissimilarity. This new equation can then be applied to tree construction, and several successful applications taking into account among-site rate variation in ribosomal RNAs have been described elsewhere (e.g. 10,17).

For the present study, 77 complete LSU rRNA sequences were analyzed, listed in Table 1. Duplicate sequences belonging to the same species were omitted from the analysis. The color map for eukaryotic LSU rRNA presented in this study can also be consulted on-line at URL http://bioc-www.uia.ac.be/u/yvdp/ . Color maps of bacterial 5S rRNA, bacterial SSU rRNA, bacterial LSU rRNA and eukaryotic SSU rRNA, published previously (1,2), can be found there as well. Secondary structures were drawn with the software tool RNAViz (18).

RESULTS AND DISCUSSION

Variability map of eukaryotic LSU rRNA

The substitution rate spectrum obtained for eukaryotic LSU rRNA is shown in Figure 1. A different color is assigned to each of the different subsets. The relative rate limits of the subsets, and the corresponding colors used in the variability map, are as follows:

$0 < v_i < 10^{-0.925}$ (blue); $10^{-0.925} < v_i < 10^{-0.425}$ (green); $10^{-0.425} < v_i < 10^{+0.075}$ (yellow); $10^{+0.075} < v_i < 10^{+0.575}$ (orange); $v_i \geq 10^{+0.575}$ (red). Since the rate distribution is not rectangular (Fig. 1), some colors are more abundant than others.

Figure 2 shows the secondary structure model of the LSU rRNA of the yeast *Saccharomyces cerevisiae* while Figure 3 shows the variability of the nucleotide sites of LSU rRNA mapped in the same shape. Colors attributed to different sites are as described above. Absolutely conserved (invariant) positions ($v_i = 0$) are indicated in purple while sites colored in pink belong to areas that are very variable, but that are deleted in too many sequences to allow a sufficiently accurate measurement of their relative evolutionary rate. The color map for LSU rRNA gives a much more detailed and quantitative description of positional variability than the crude distinction between variable and conserved areas that is often made by visual inspection of sequence alignments.

It can be seen that in general the two nucleotides of a base pair have the same or a neighboring color, i.e. they are about equally variable. This is as expected, since the substitution of a base-paired nucleotide generally requires a compensating substitution in the opposite strand. Exceptions are mostly due to the fact that in some cases a particular base, usually a G or a U, seems to be required in one strand, but the existence of G·U pairs aside from G·C and A·U pairs allows the complementary base to change more freely.

As can be seen in Figure 3, 10 highly variable areas (orange, red and pink) can be distinguished in the eukaryotic LSU

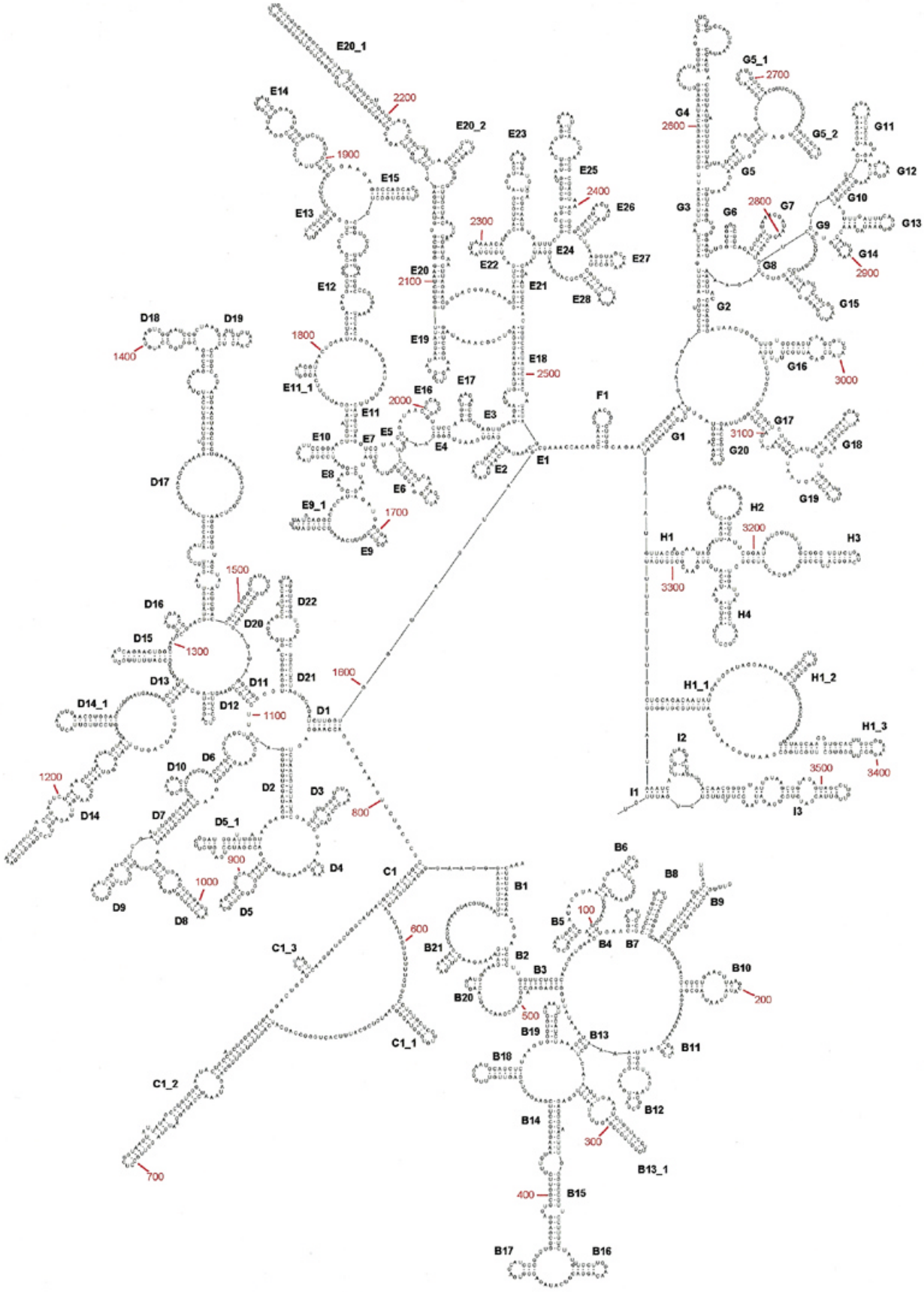


Figure 2. Secondary structure model for the LSU rRNA of *S.cerevisiae*. The sequence is written clockwise from 5' to 3', sites are numbered in red every 100 nucleotides. Helix numbering is according to De Rijk *et al*. (16).

Table 2. Fraction of more conserved and less conserved sites in double-stranded areas and in different types of loops

Relative substitution rate v_i	No. of sites in different structural elements					
	multi-branched loops	hairpin loops	internal- and bulge loops[a]	all single stranded elements	helices	entire molecule[b]
$v_i < 10^{-0.925}$	252	180	181	613	548	1161
$v_i \geq 10^{-0.925}$	322	201	268	791	1124	1915
Total	574	381	449	1404	1672	3076
% with $v_i < 10^{-0.925}$	43.9	47.2	40.3	43.7	32.8	37.7
% with $v_i \geq 10^{-0.925}$	56.1	52.8	59.7	56.3	67.2	62.3

[a]The number of sites in bulge loops was too small (43 sites) to make a separate calculation meaningful.
[b]This includes all sites for which a relative substitution rate was computed, i.e. all but those deleted in >75% of the sequences and indicated in pink on the color map of Figure 3.

rRNA. These are formed by the following helices: B8; B13_1 to B16; the whole area denoted as C; D5 and D5_1; D14_1; D20; E20_1 to E20_2; G5_1 to G5_2; and H1_1 to H1_3. Many of the variable areas are characterized by major size variations. For example, the areas enclosed by helices C1 and E20, the entire area H1_n, and to a lesser extent helix D20, are also hot-spots for extremely variable insertions. These insertions were first described by Hassouna *et al*. (19) who referred to them as D(ivergent)-domains. As a rule, strong length heterogeneity seems to be most common in apical helices, i.e. those ending in a hairpin loop. Helices formed by long distance interactions, i.e. those bounded by multibranched loops, have less freedom to change in length. It should also be noted that the LSU rRNA molecule contains a number of potential branching points that bear additional helices in a limited set of species. For example, helices B14 and B15, though separated only by an internal loop in *S.cerevisiae*, were numbered differently because a potential branching point separates them.

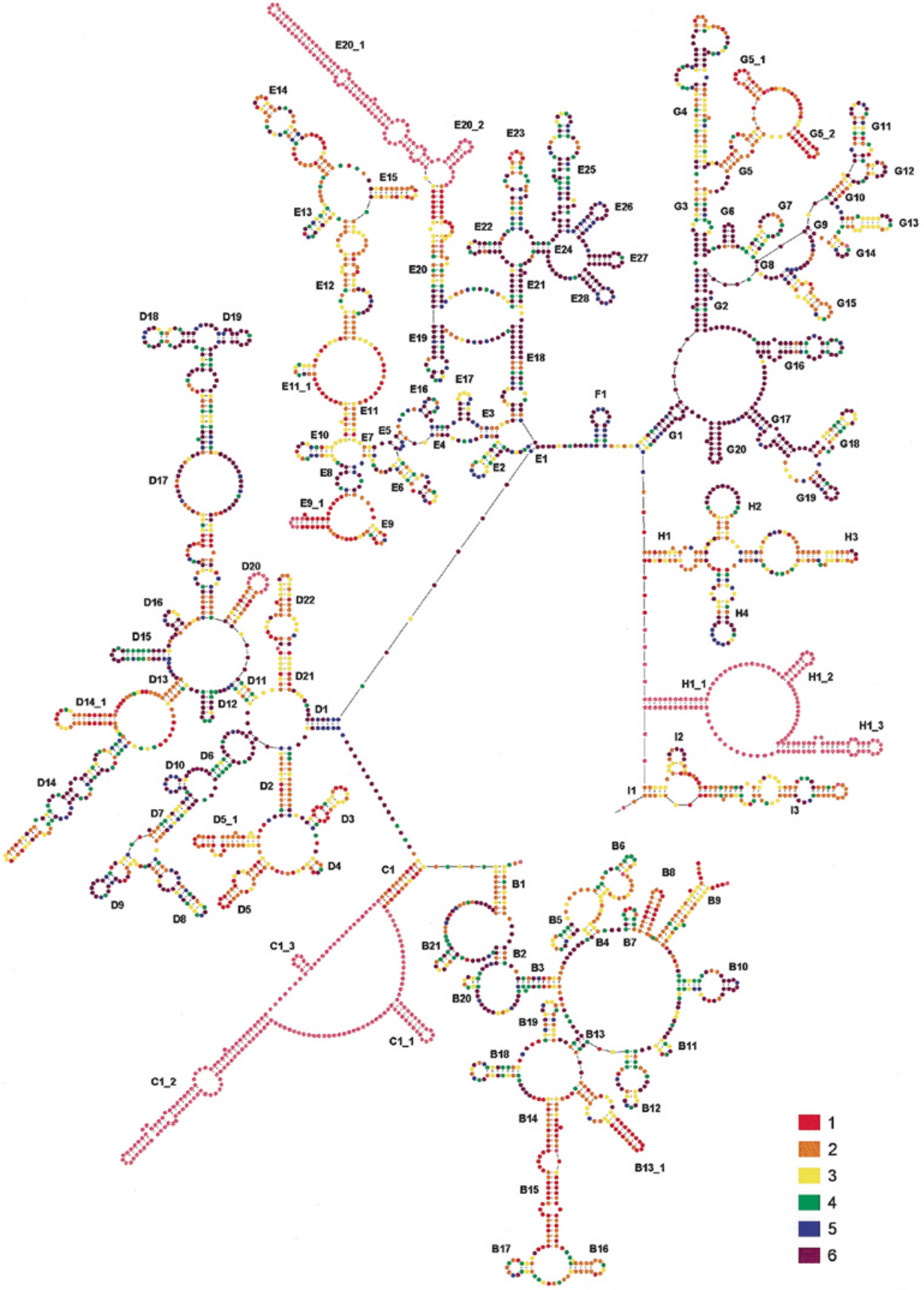
Beside highly variable regions, several regions of conserved nature can be distinguished in the LSU rRNA. For the LSU rRNA of prokaryotes and for the SSU rRNA of prokaryotes and eukaryotes a stronger sequence conservation in single-strands than in double-strands has been reported (1,8,20–22). In order to examine this quantitatively in the case of eukaryotic LSU rRNA, separate substitution rate spectra were measured for nucleotide sites involved in base pairing and for those forming part of each type of single-stranded structural elements: multibranched-, hairpin-, internal- and bulge-loops. These spectra (not shown) were not very different from that measured for the entire molecule (Fig. 1). However, it should be remembered that the spectra show only the distribution of sites with a measurable substitution rate and do not include sites that are identical in all hitherto sequenced molecules. The latter have $v_i = 0$ and therefore cannot be represented on a logarithmic scale. In order to demonstrate the greater tendency of conservation in single-stranded areas, sites were divided arbitrarily in more conserved ones ($v_i < 10^{-0.925}$, i.e. purple and blue in Figs 1 and 3) and more variable ones ($v_i \geq 10^{-0.925}$, i.e. green to red in Figs 1 and 3). Table 2 shows that the fraction of more conserved sites is larger in single-stranded structural elements than in helices. In Figure 4 the fraction of more conserved sites ($v_i < 10^{-0.925}$) is calculated in both single- and double-stranded regions over the entire molecule in a sliding window of 50 nucleotides. In this calculation positions that contain a nucleotide in <25% of the sequences (pink in Fig. 3) were included with the most variable ones ($v_i \geq 10^{-0.925}$). This graph shows more clearly that, overall, single-stranded regions have a higher fraction of conserved positions than double-stranded regions of the LSU rRNA molecule. One notable exception is the area between positions 2000 and 2400, which corresponds approximately to the region covered by helices E16–E25.

Very often, specific functions can be ascribed to regions of the molecule that are conserved in structure (e.g. 23,24). One of the most conserved areas, both in bacterial (1) and eukaryotic LSU rRNA is the large multibranched loop in area G and the helices surrounding it (Fig. 3). This structure is generally considered to be the major element of the peptidyl transferase center of the ribosome, which is the catalytic center responsible for the peptide bond formation (25,26). Other highly conserved structures in the LSU rRNA are helices D18–D19 which are part of the so-called GTPase center of the ribosome, helices E21–E28 and the hairpin loop of helix H2 (25).

ACKNOWLEDGEMENTS

We want to thank Linda Medlin for DNA of haptophytes. Our research is supported by the Special Research Fund of the University of Antwerp and by the Fund for Scientific Research, Flanders. Yves Van de Peer is Research Fellow of the Fund for Scientific Research, Flanders.

Figure 3. (Next page) Variability map superimposed on the LSU rRNA secondary structure model of *S.cerevisiae*. Nucleotides are subdivided into five groups of increasing variability (see text for details). The most variable positions are in red, the least variable ones in blue. Absolutely conserved positions in all structures hitherto known are indicated in purple. Hypervariable regions that were not taken into consideration for rate calibration, because they are absent in >75% of the eukaryotic sequences considered, are indicated in pink. These include C1_1 to C1_3; E20_1 to E20_2; H1_1 to H1_3, the hairpin loops of D20 and E9_1, as well as individual nucleotides peculiar to the *S.cerevisiae* LSU rRNA.


E20_1
E14
E20_2
E23
E25
E15
E26
E13
E22
E24
E27
E20
E21
E12
E28
D18
D19
E19
E18
E11_1
E17
E16
F1
E11
E3
E10
E7
E5
E4
G1
E1
E8
E2
E6
D17
E9_1
E9
D20
D22
D16
D15
D21
D13
D11
D14_1
D12
D1
D6
D10
D2
D14
D7
D5_1
D3
D4
D9
D8
D5
C1
B1
C1_3
B21
B2
B3
B20
C1_1
B19
B13
C1_2
B18
B14
B15
B17
B16
B6
B8
B9
B5
B4
B7
B10
B11
B12
B13_1
G4
G5_1
G5_2
G11
G5
G12
G3
G6
G7
G10
G9
G13
G14
G8
G15
G2
G16
G17
G18
G20
G19
H2
H1
H3
H4
H1_1
H1_2
H1_3
I2
I1
I3
1
2
3
4
5
6

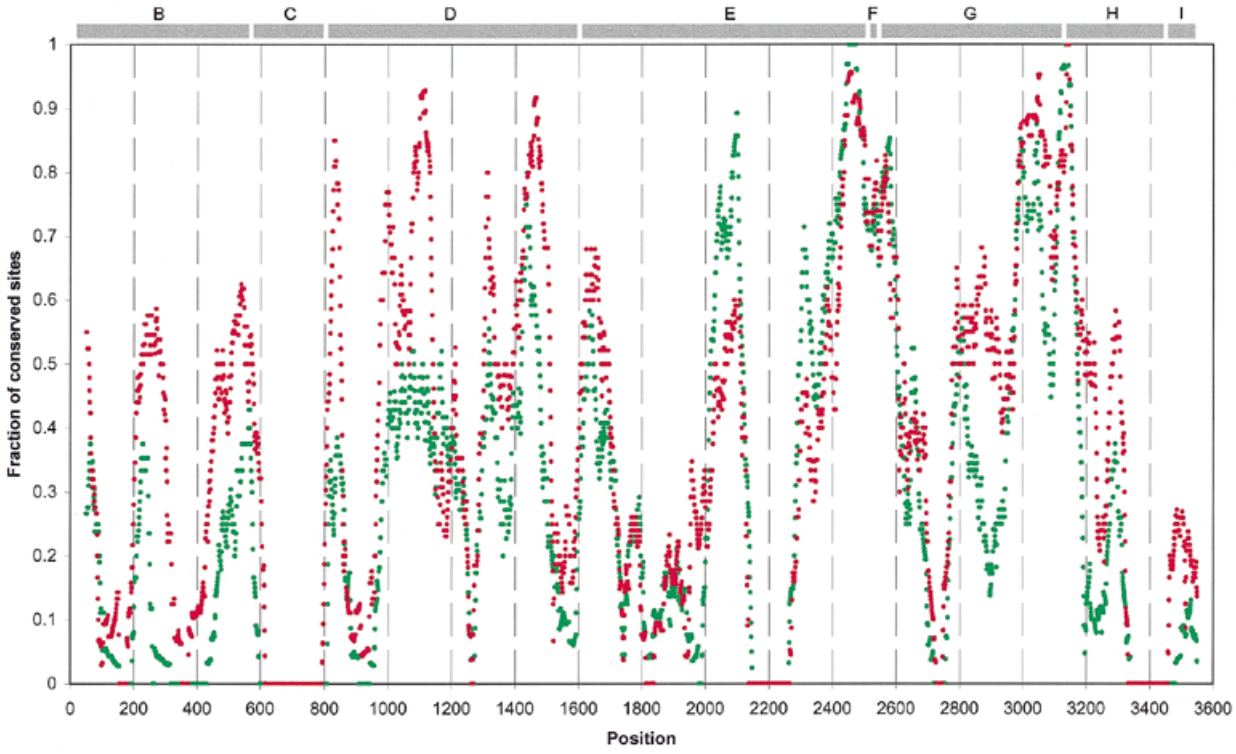


Figure 4. Graph showing the fraction of conserved sites, counted with a sliding window of 50 nucleotides. Each red dot represents the fraction of positions with $v_i < 10^{-0.925}$ in single stranded sites, each green dot represents this fraction in helical sites. The location of structural areas B to I is indicated by bars on top of the figure. Long areas with a fraction of conserved sites equal to 0 around positions 700, 2200 and 3400 consist of sites of presumably high but unmeasurable variability (pink dots in Fig. 3).

REFERENCES

1. Van de Peer,Y., Chapelle,S. and De Wachter,R. (1996) *Nucleic Acids Res.*, **24**, 3381–3391.
2. Van de Peer,Y., Jansen,J., De Rijk,P. and De Wachter,R. (1997) *Nucleic Acids Res.*, **25**, 111–116.
3. Van de Peer,Y., Neefs,J.-M., De Rijk,P. and De Wachter,R. (1993) *J. Mol. Evol.*, **37**, 221–232.
4. Van de Peer,Y., Van der Auwera,G. and De Wachter,R. (1996) *J. Mol. Evol.*, **42**, 201–210.
5. Van der Auwera,G. and De Wachter,R. (1997) *J. Mol. Evol.*, **45**, 84–90.
6. Van der Auwera,G., Hofmann,C.J., De Rijk,P. and De Wachter,R. (1998) *Mol. Phyl. Evol.*, **10**, 333–342.
7. Knoll,A.H. (1992) *Science*, **256**, 622–627.
8. Abouheif,E., Zardoya,R. and Meyer,A. (1998) *J. Mol. Evol.*, **47**, 394–405.
9. Olsen,G.J. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **LII**, 825–837.
10. Van de Peer,Y., Rensing,S., Maier,U.-G. and De Wachter,R. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 7732–7736.
11. Yang,Z. (1996) *Trends Ecol. Evol.*, **11**, 367–372.
12. Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.*, **21**, 3055–3074.
13. De Rijk,P., Van de Peer,Y., Chapelle,S. and De Wachter,R. (1994) *Nucleic Acids Res.*, **22**, 3495–3501.
14. De Rijk,P., Van de Peer,Y. and De Wachter,R. (1996) *Nucleic Acids Res.* **24**, 92–97.
15. De Rijk,P. and De Wachter,R. (1993) *Comput. Appl. Biosci.*, **9**, 735–740.
16. De Rijk,P., Robbrecht,E., de Hoog,S., Caers,A., Van de Peer,Y., De Wachter,R. (1999) *Nucleic Acids Res.*, **27**, 174–178.
17. Van de Peer,Y. and De Wachter,R. (1997) *J. Mol. Evol.*, **45**, 619–630.
18. De Rijk,P. and De Wachter,R. (1997) *Nucleic Acids Res.*, **25**, 4679–4684.
19. Hassouna,N., Michot,B. and Bachellerie,J.-P. (1984) *Nucleic Acids Res.*, **12**, 3563–3581.
20. Vawter,L. and Brown,W.M. (1993) *Genetics*, **134**, 597–608.
21. Rzhetsky,A. (1995) *Genetics*, **141**, 771–783.
22. Otsuka,J., Terai,G. and Nakano,T. (1999) *J. Mol. Evol.*, **48**, 218–235.
23. Egebjerg,J., Larsen,N. and Garrett,R.A. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution*. American Society of Microbiology, Washington, DC, pp. 168–179.
24. Noller,H.F., Moazed,D., Stern,S., Powers,T., Allen,P.N., Robertson,J.M., Weiser,B. and Triman,K. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution*. American Society of Microbiology, Washington, DC. pp. 73–92.
25. Raué,H.A., Musters,W., Rutgers,C.A., van't Riet,J. and Planta,R.J. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds.) *The Ribosome. Structure, Function and Evolution*. American Society of Microbiology, Washington, DC, pp. 217–235.
26. Bocchetta,M., Xiong,L. and Mankin,A.S. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 3525–3530.